Analyzing Sentiment Towards the Covid-19 Vaccine via Twitter
Stacy Fortes and Daehan Kwak
Kean University, School of Computer Science and Technology, Union, NJ 07083

Keywords: Covid-19, Vaccine, Vaccine Hesitancy, Sentiment Analysis, Twitter API

Since March of 2020, when the public began quarantining due to Covid-19, people have been waiting for life to go back to normal. Up until April of 2021, "normal" activities were only allowed by being socially distant or wearing a mask. However, in early 2021, the director of the CDC announced that a vaccine would soon be released to the public. For some, this announcement seemed to be good news, but for others, this was another obstacle in the way to normalcy.

For years, vaccines have been recognized as the most successful measure in preventing massive outbreaks of viruses such as Covid-19. However, some people have shown pushback against the Covid-19 vaccine due to the uncertainty of its effectiveness coupled with its potential side effects. Vaccine hesitancy not only has a negative impact on society but also poses a real threat to public health [1]. This is an important issue around the world and questions arise on how different states in the United States feel about getting vaccinated. Therefore, the purpose of this study is to analyze the sentiment of Twitter users towards vaccination, specifically the Covid-19 vaccine in the United States.

This study collects data through a Twitter mining system to pick up on hashtags and keywords. The primary language used is Python along with the use of a Twitter API, Tweepy [2]. The hashtags and keywords used are the following: "covid vaccine", "moderna", "pfizer", "#CovidVaccine", "johnson & johnson". The tweets are then saved to a CSV file and filtered by location in order to remove any locations outside of the United States. Once filtered, the data is then processed through a sentiment analysis API from the Natural Language Toolkit (NLTK) called Vader. Vader uses a list of lexical features in order to label a piece of text as positive, negative, or neutral by giving numerical values to each category [3]. The numerical results of sentiment analysis per state are then plugged into a spreadsheet to create a visualization. The visualization is a color-coded map of the USA in which green indicates the most positive sentiment and red indicates the most negative.

Initially 5000 tweets were collected. However, in order to get a more accurate and all-encompassing representation of sentiment towards the Covid-19 vaccine in the United States, we determined that more data was needed. Due to limitations of data collection through Tweepy, we turned to an existing dataset of which contained nearly 150,000 tweets spanning from March 2020 to August 2021 [4]. Because this dataset consists of general tweets, it is cleaned up to contain only the appropriate data for this study. This is done by removing all tweets from locations outside of the country and then filtering the tweets to only include those which contained one of the new keywords such as: "vaxx", "vaccine", "#vaccine", "#coronavaccine", "Pfizer", "BioNTech", "Moderna", "AstraZeneca", "AZ", "Janssen", etc. The reason for updating the keyword list is to include more Covid-19 vaccine brands. This larger dataset is processed through a sentiment analysis tool in order to obtain the sentiment of all 50 states. These sentiment results are then averaged and visualized on a map. This dataset will also be processed

through a topic modeling [5] in order to analyze where exactly the Covid-19 vaccine is being discussed most frequently in the United States. From there we then further analyze whether this conversation is positive or negative. This analysis will produce a more comprehensive understanding of sentiment towards the vaccine around the United States.

Covid-19 has been plaguing the world for almost two years now and one of the only ways to return to normal is to obtain herd immunity through vaccination [6]. Vaccine hesitancy is a real and prevalent issue in our society which must be addressed. This research can be used to inform health officials about the general sentiment toward the vaccine in each state and provide insight as to which states need to be further educated on the benefits of the vaccine. Additionally, this research can be compared to live Covid-19 vaccination rates in the future to observe whether a state expressing negative sentiment on Twitter correlates to a lower vaccination rate in real life or vice versa. By analyzing the sentiment towards the Covid-19 vaccine we hope to provide the first step towards mitigating the risk associated with vaccine hesitancy.

References
[1] Daniel A. Salmon, Matthew Z. Dudley, Jason M. Glanz, Saad B. Omer, "Vaccine Hesitancy: Causes, Consequences, and a Call to Action", American Journal of Preventive Medicine, Volume 49, Issue 6, Supplement 4, 2015, Pages S391-S398, ISSN 0749-3797, https://doi.org/10.1016/j.amepre.2015.06.009.
[2] "Tweepy Documentation," Tweepy Documentation - Tweepy 3.10.0 Documentation, docs.tweepy.org/en/stable/.
[3] "Nltk.sentiment Package," Nltk.sentiment Package - NLTK 3.6.3 Documentation, www.nltk.org/api/nltk.sentiment.html.
[4] Rabindra Lamsal, April 26, 2020, "Coronavirus (COVID-19) Geo-tagged Tweets Dataset", IEEE Dataport, doi: https://dx.doi.org/10.21227/fpsb-jz61.
[5] Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kim, K. and Kwak, K. S. (2019). Transportation sentiment analysis using word embedding and ontology-based topic modeling. Knowledge-Based Systems, 174, 27-42. doi: https://doi.org/10.1016/j.knosys.2019.02.033
[6] Mayo Foundation for Medical Education and Research. (2021, August 28), "Herd immunity and covid-19 (coronavirus): What you need to know", Mayo Clinic, https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/herd-immunity-and-coronavirus/art-20486808.